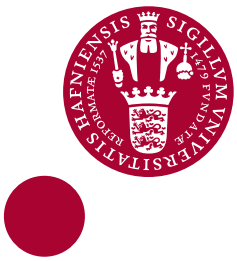

Influence of packaging and storage time on aroma compounds of minimally processed lettuce

Deza Durand, Karla Michelle; Petersen, Mikael Agerlin; Roepstorff, Allan Knud; Poll, Leif

Published in:
Advantages and challenges in flavor chemistry and biology

Publication date:
2011

Document version
Publisher's PDF, also known as Version of record

NORMALIZATION IS A NECESSARY STEP IN NMR DATA PROCESSING: FINDING THE RIGHT SCALING FACTORS

F. Capozzi[1,2], A. Ciampa[1], G. Picone[1], G. Placucci[1] and F. Savorani[3]

[1]Department of Food Science, University of Bologna, Piazza Goidanich 60, I-47521 Cesena, Italy, Email: francesco.capozzi@unibo.it
[2]CIRMMP, Via L. Sacconi, 6, I-50019 Sesto Fiorentino (FI), Italy
[3]Department of Food Science, Quality & Technology, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark

1 INTRODUCTION

Metabolites are the end products of cellular regulatory processes, and their levels can be regarded as the ultimate response of biological systems to genetic or environmental changes. In parallel to the terms 'transcriptome' and 'proteome', the whole set of metabolites synthesized by a biological system constitutes its 'metabolome'.[1]

Conventionally, the analytical procedure can be restricted to the identification and quantification of a selected number of pre-defined metabolites in a biological sample. This process is called metabolite profiling (or, sometimes, metabolic profiling).[2] For example, pre-defined metabolites may belong to a class of compounds (such as polar lipids, isoprenoids, or carbohydrates), or be narrowed down to members of particular pathways. The term metabolite profiling is frequently used in the specific context of drug research in the description of catabolic degradation of an applied chemical.

Conversely, the approach revealing the comprehensive metabolome of the biological system under investigation should be called metabolomics. Metabonomics is a further specialization of the *omics* approach, which is based on the systematic profiling of metabolite levels, but encompasses the systematic and temporal changes of the metabolome in whole organisms occurring in response to diet, lifestyle, environment, etc.[3] Metabonomics must aim at avoiding exclusion of any metabolite by using well conceived sample preparation procedures and analytical techniques.

The resolving power of the chosen analytical method must be high enough to maintain sensitivity, selectivity, matrix independence, and universal applicability. Since data sets for metabonomics are complex by nature, adequate tools are needed to handle, store, normalize, and evaluate the acquired data in order to describe the systemic response of the biological system. Some tools can be adopted by classical structure analysis, such as mass spectrometry (MS) and nuclear magnetic resonance (NMR).[4]

NMR and MS are mostly used to generate global metabolite profiles in metabolomics.[5,6,7] An advantage of NMR is that the samples normally do not require any physical or chemical treatment prior to the analysis, whereas MS usually requires that the metabolites be separated from the sample before detection by liquid chromatography.[8] Although MS is more sensitive than NMR, the latter is a more attractive tool for metabonomics studies because the non-destructive nature of NMR enables observation of

the dynamics as well as separation of metabolites in biological samples; in contrast, MS disrupts the structures and the interactions of molecular complexes. Chemometrics, e.g. NMR spectroscopic analysis coupled with multivariate statistical methods, offer a powerful new approach for assessing metabolic function.[9,10]

Pattern recognition and related multivariate statistical approaches can be used to discern significant patterns in complex data sets with the aim of classifying objects by identifying inherent patterns in a set of indirect measurements. Pattern recognition methods, such as principal components analysis (PCA) and classification methods such as partial least-squares discriminant analysis (PLS-DA) and orthogonal projection to latent structures discriminant analysis (OPLS-DA), reduce the dimensionality of complex data sets and thereby facilitate the visualization of inherent patterns in the data accelerating the interpretation.

^{1}H NMR-based metabonomics has been applied in the food sciences,[11] including assessments of green tea,[12] rosemary,[13] honey,[14] grape wine,[15,16] crops,[17,18] etc. Mainly, the spectroscopic data ^{1}H-NMR are analyzed by multivariate data analysis methods to extract information about changes in distinct metabolites of examined samples.

The steps involved in the analysis of metabonomics are as follows:[19,20] (a) post-instrumental processing of acquired spectroscopic data, such as removal of offsets by polynomial baseline correction, calculation of intensity values either on each data point, on each peak, or summed over segmented regions (binning); (b) production of a data table from the analytical measurements such that there are m rows (observations, samples) and n columns (variables, frequencies, integrals); (c) normalization of the data or some related adjustment to the spectral intensities (a row operation); (d) scaling of the data (a column operation); (e) multivariate statistical modelling of the data.

The normalization step can is applied to the data from each sample and comprises methods to make the data from all samples directly comparable with each other.[20,21] One of its common applications is to remove or minimize the effects of variable dilution of the samples. Dilution is defined as a process influencing the concentrations of all metabolites, and thus all peak intensities of the corresponding spectrum, by the same factor (coefficient), which can also be referred to as unspecific changes of metabolites.

In contrast, metabolic responses influence a few metabolites in the sample and, consequently, only a few peaks of the corresponding spectrum. These specific changes are visible as relative changes of concentrations of few metabolites related to the concentrations of all other metabolites, which represent the overall concentration of the sample. Usually these specific relative changes are of interest in metabonomics studies in contrast to the overall concentration of the sample, and are only visible after an adequate normalization of the spectra.

Normalization can also be necessary due to technical reasons. If spectra are recorded using a different number of scans, or if spectra are recorded with different devices, the absolute values of the spectra are different and rendering a joint analysis of spectra without prior normalization is impossible.

The specific object of the present investigation was to find a new algorithm, developed on a dataset of NMR spectra recorded on 20 mixtures of 10 different metabolites, also exerting signal overlap, able to correct the dilution errors that have been purposely and artificially introduced during sample preparation. The developed algorithm was also applied on a "natural" spectral dataset of tomatoes extracts, for which both NMR and HPLC techniques are applied in a non-hyphenated manner.

2 METHOD AND RESULTS

2.1 Samples preparation

2.1.1 The mixtures. Twenty mixtures, each containing glutamic acid, p-OH-benzoic acid, histidine, proline, valine, serine, imidazole, aspartic acid, adenosine 5'-triphosphate, phenylalanine, were prepared in 50 mM phosphate buffer at pH 7, by mixing 10 µl of 50 mM solutions of the first 5 substances, and 5 to 25 µl of 50 mM solutions of the remaining five chemicals. According to this procedure, half metabolites were initially kept at constant concentration in all mixtures, whereas the concentration of the other half metabolites was allowed to vary from 0.5 to 2.5 times the amount of those kept at constant concentration. Subsequently, all mixtures were diluted adding different known volumes of water, from 0.87 to 2.80 ml. From each mixture 800 µl were transferred in a NMR tube for the spectroscopic analysis.

2.1.2 Cherry tomatoes. Eleven different samples of cherry tomatoes have been grinded, freeze-dried and stored under argon atmosphere. Before NMR analysis, 2 g of freeze-dried powder have been extracted with 50 ml of $CHCl_3$, for 48 hours, under magnetic stirring. The suspension has been filtered and concentrated to 1 ml, by solvent evaporation under N_2 flow. A 0.1 ml aliquot of this sample has been diluted to 0.5 ml with $CHCl_3$, for the HPLC analysis, whereas the remaining 0.9 ml aliquot has been completely dried and re-solubilised in 1 ml of $CDCl_3$ 99,8 % + TMS 0.05% for the NMR analysis.

2.2 NMR acquisition parameters

2.2.1 Mixtures samples. 1D ^{1}H- NMR spectra were recorded at 298 K on a Varian Mercury Plus AS400 spectrometer operating at 400.097 MHz. The pulse sequence used contains the presaturation step of the water signal, obtained by centring the spectral window at 4.706 ppm and using a 2 s pulse with an attenuation of 5 dB. The spectral window (SW) was 14 ppm, the number of points of the spectrum were 1 K and 90° proton pulses (9.30 µs at 55 dB attenuation) with 6 s time delay were used. Each spectrum was obtained over 512 scans and the FID, prior to Fourier transformation, was multiplied by an exponential factor (line broadening) equal to 1.5 Hz. Phase and baseline corrections of spectra were manually performed by using the software package MestRe-C 4.9.8.0 (www.mestrec.com, MestRe-C Research SL, Santiago de Compostela, Spain). This software was also used to operate integration of each spectrum by selecting 30 integral regions and to export data in ASCII. Afterwards data were statistically processed by using the free software package *R* version 2.11.1.

2D ^{1}H,^{1}H TOCSY was registered with the standard "mlevphpr" pulse program, with water presaturation during relaxation delay, a spectral width of 6 kHz in both dimensions, a 6 s relaxation delay, and 80 ms mixing time, 1 K data points in F2 and 512 increments in F1. Also in this case, phasing of the two-dimensional spectrum was performed by using the software package MestRe-C.

2.2.2 Cherry tomatoes samples. All 1D ^{1}H-NMR spectra have been recorded using a Varian Mercury Plus AS400 spectrometer, operating at 400.097 MHz ^{1}H Larmor frequency. 1D ^{1}H-NMR experiments have been recorded at 298 K with a 30° pulse (2.1 µs) over a spectral width of 5583.47 Hz (14 ppm), 16K complex data points, 1s relaxation delay, 1024 scans, 20 Hz spinning. Each spectrum has been Fourier transformed using a 0.5 Hz Line Broadening apodization. The phase and baseline correction as well as the chemical shift calibration have been accurately performed over all the spectra by using the software MestRe-C. The same program has been used to export data in ASCII format with

16K data points. Chemical shift referencing has been performed by setting at 7.26 ppm the signal of the residual $CHCl_3$ solvent in all spectra.

A single 2D ^{1}H-^{1}HTOCSY spectrum has been acquired using a Bruker spectrometer operating at 800.1338 MHz equipped with a cryoprobe, under the following conditions: 298°K, TPPI phase sensitive mode, the standard mlevphpr pulse program, 0.213 s of acquisition time, 1K data complex points in F2 and 512 increments in F1, spectral width of 9615.38 Hz, 80 ms mixing time and a relaxation delay of 2.5 s. In order to prevent errors due to sample modifications during storage, a 1D ^{1}H-NMR spectrum of the same sample has been acquired before and after the TOCSY experiment. A comparison of the 1D spectra pre and post TOCSY experiment confirmed that no appreciable molecular modifications happened during the acquisition time. Also the 2D TOCSY spectrum has been processed with MestRe-C.

2.3 HPLC analysis

HPLC analyses have been carried out using an HP 1100 series instrument (Agilent Technologies, Palo Alto, CA), equipped with a binary pump delivery system, a degasser, an autosampler, and an HP diode array UV-vis detector (DAD). Chromatography has been performed with a reverse-phase Luna C18 (Phenomenex, Torrance, CA) column of 5 μm particle size and 25 cm x 3.00 mm i.d. The mobile phase flow rate has been chosen in 0.7 $ml \cdot min^{-1}$. The acquisition wavelength has been set at 455 nm. The injection volume was 10 μL, which is an aliquot withdrawn from the same samples utilized for the NMR analysis. All of the analyses have been carried out at 303 °K using a thermostatic oven. The gradient elution has been performed using a mobile phase of MetOH/H_2O/ACN (80/10/10) pump A; ACN pump B. Gradient profile: from 0% B up to 95% in 60 min, 95% B for 59 min, down to 0% B in 10 min for a total of 129 min of elution time. All the HPLC chromatograms have been processed with LC/MSD ChemStation Rev. A.08.03 (Agilent Technologies) in order to perform a blank subtraction and a baseline correction. The same program has been used to export chromatograms in ASCII format with 19350 data points each one.

2.4 The TOCSY-filter normalization algorithm

The normalization algorithm, that has been developed and tested within this study, executes the following steps (*in italics*) in the R-project programming environment:
1) Chose the cross-correlations to be included in the optimization process (3 false and 1 true)
2) Define the first false cross-correlation as the one between integral regions #16 and #21
```
a1f <- 16; b1f <- 21
```
3) Define the second false cross-correlation as the one between integral regions #21 and #27
```
a2f <- 21; b2f <- 27
```
4) Define the third false cross-correlation as the one between integral regions #20 and #30
```
a3f <- 20; b3f <- 30
```
5) Define the true cross-peak as the one between integral regions #14 and #15
```
aV <- 14; bV <- 15
```
6) `'dil'` represents the vector of 20 dilution coefficients, initially set equal to 1
```
dil <- rep(1,20)
```
7) Define the optimization function `'crosscorr'(`*`'integrals'`* represents the non-normalized NMR integrals data matrix)

```
Crosscorr <- function(dil)
{p1 <- integrals[,a1f] * dil
q1 <- integrals[,b1f] * dil
regress.f1 <- lm(formula = p1~q1)
r1f <- summary.lm(regress.f1)$r.squared
p2 <- integrals[,a2f] * dil
q2 <- integrals[,b2f] * dil
regress.f2 <- lm(formula = p2~q2)
r2f <- summary.lm(regress.f2)$r.squared
p3 <- integrals[,a3f] * dil
q3 <- integrals[,b3f] * dil
regress.f3 <- lm(formula = p3~q3)
r3f <- summary.lm(regress.f3)$r.squared
pV <- integrals[,aV] * dil
qV <- integrals[,bV] * dil
regress.f6 <- lm(formula = pV~qV)
rV <- summary.lm(regress.f6)$r.squared
filter = r1f + r2f + r3f - rV
return(filter)}
```

'*filter*' is the objective function obtained by combination of the three false (rNf) and one true (rV) cross-correlation elements characterized by their regression coefficients, R^2, and undergoing minimization through the optimization of the '*dil*' vector.

```
8)res<-optim(dil, crosscorr, control=list(maxit=20000))
```

'*res$par*' is the vector containing the calculated dilution coefficients returned from the algorithm.

2.5 The TOCSY-filter normalization algorithm for ^{1}H-NMR metabonomics

The effect of the variable large amount of water in food samples is of paramount importance in metabonomics study, since the consequent dilution errors concerning all metabolites dissolved in their extracts affect the proper data processing. For this reason, the samples are often pre-treated for preservation purposes, but not only, through freeze-drying with quantitative elimination of water, which in most cases, being the main component, can be as high as 90% of weight. An inconsistent and unreliable freeze-drying process, or a bad conservation easily lead to a not reproducible application of the normalization step, and then to a wrong and inconsistent pre-processing of data, which can vary even among different replicates of the same sample.

2.5.1 Evaluation of the normalization algorithm on an artificial set of mixtures. In order to develop an algorithm able to correct these dilutions errors on real matrices, at first data were collected on ^{1}H-NMR spectra acquired on an artificial trial set of twenty mixtures, each one consisting of variable concentrations of the same ten molecules. In each mixture, five substances were always kept at a constant concentration (glutamic acid, p-OH-benzoic acid, histidine, proline, valine), while the remaining five chemicals (serine, imidazole, aspartic acid, adenosine 5'-triphosphate, phenylalanine) were added at different concentrations. The mixtures were then further diluted with appropriate additions of H_2O in order to have dilution coefficients, K, ranging between the maximum value of 5.8, for the more diluted mixture, and the minimum value of 2.0, for the more concentrated one.

The ^{1}H-NMR spectra were calibrated with respect to the chemical shift of the valine methyl signal falling at 0.99 ppm. In Table 1, the 36 signals assigned in the ^{1}H-NMR spectra of the trial mixtures are listed, giving rise to 30 integral regions generating a 20 rows/spectra x 30 columns/integrals data matrix. Six signals (namely 1, 5, 12, 13, 26, 28)

are indeed overlapped to other six resonances (1', 5', 12', 13', 26', 28', respectively), and each pair is included in the same integral region.

Table 1 *Resonance assignments with chemical shift and multiplicity of the signals from all metabolites present in the 30 integral regions selected in the ^{1}H-NMR spectra (s – singlet, d – doublet, t – triplet, m – multiplet, dd – double doublet, dt – double triplet). Chemical shifts with the same signal number (still distinguished by the prime sign) refer to signals belonging to different metabolites but falling in the same integral region.*

Signals	^{1}H Shift (ppm)	Molecule	Assigment	Multiplicity
1	8,357	adenosine 5'-triphosphate	2'-CH	s
1'	8,326	imidazole	C2-H ring	s
2	8,107	adenosine 5'-triphosphate	8'-CH	s
3	7,997	histidine	C2-H ring	s
4	7,638	p-OH-benzoic acid	α-CH	d
5	7,231	phenylalanine	C2-H ring	m
5'	7,153	imidazole	C4-H ring	s
6	7,012	histidine	C4-Hring	s
7	6,760	p-OH-benzoic acid	β-CH	d
8	5,963	adenosine 5'-triphosphate	1' ribose	m
9	4,414	adenosine 5'-triphosphate	2' ribose	m
10	4,227	adenosine 5'-triphosphate	3'-4' ribose	m
11	4,070	adenosine 5'-triphosphate	5' ribose	m
12	3,961	proline	α-CH	t
12'	3,945	phenylalanine	α-CH	dd
13	3,835	histidine	α –CH	dd
13'	3,804	serine	β –CH	dd
14	3,726	glutamic acid	α-CH	t
15	3,679	serine	α-CH	dd
16	3,585	aspartic acid	α-CH	dd
17	3,444	valine	α-CH	d
18	3,210	proline	δ-CH	t
19	3,163	proline	δ'-CH	t
20	3,100	phenylalanine	β-CH	dd
21	3,037	histidine	β-CH	dd
22	2,959	histidine	β'-CH	dd
23	2,662	glutammic acid	β-CH	dt
24	2,615	glutammic acid	β'-CH	dt
25	2,521	glutammic acid	γ-CH	t
26	2,177	proline	β-CH	m
26'	2,177	aspartic acid	β-CH_2	dd
27	2,099	valine	β-CH	m
28	1,848	proline	γ-CH	m
28'	1,911	aspartic acid	β'-CH_2	dd
29	1,160	impurity	?	s
30	0.990	valine	γ, γ' –CH_3	d

The aim of the comparison between the STOCSY autocorrelation matrix and the simplified TOCSY array is to verify whether only the signals belonging to the same molecule are correlated with high values of R^2 in the autocorrelation matrix.

The NMR dataset was used to generate the corresponding autocorrelation matrix, whose cross-correlation elements possess coefficients R^2 resulting from the regression analysis among all column pairs representing the 30 integral regions (Figure 1A). The autocorrelation matrix is also known as the statistical total correlation spectroscopy

analysis (STOCSY).[22] The highest correlations are represented by black and dark grey dots, with R^2 greater than 0.8 and 0.7, respectively, while the least correlated integral regions are coded in medium and light grey, with R^2 less than 0.5 and 0.4, respectively. The autocorrelation matrix resulting from the original non-normalized spectra is then compared with another array (Figure 1B) that is the simplified binned representation of the 2D-TOCSY spectrum acquired on the most concentrated mixture of metabolites (Figure 2), in a way that the two matrices dimensionally fit. The off-diagonal dots in Figure 1B represent true correlations among protons belonging to the same molecule. The TOCSY spectrum, indeed, permits to identify all the signals of an entire spin system of the molecules dissolved in the mixture. For example, by inspecting the valine signals, we can assign the vicinal proton pairs in α– and β–CH, as well as the γ, γ' methyl groups, with respect to carboxyl group carbon, that fall, respectively, to 3.44, 2.09, and 0.99 ppm.

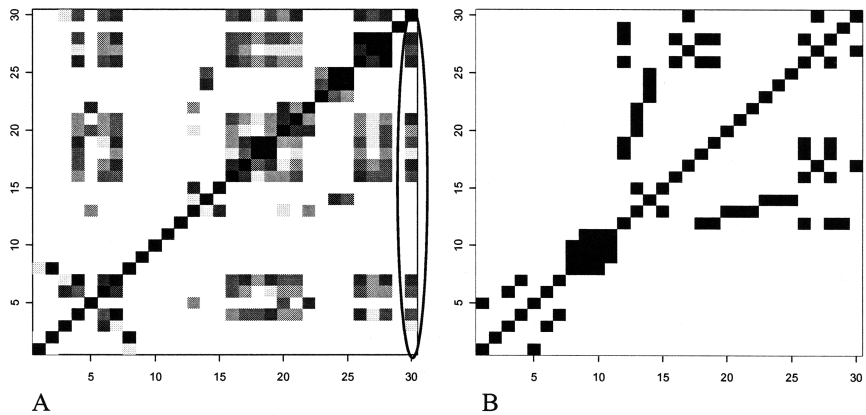

Figure 1 A) *Autocorrelation matrix obtained from the regression analysis among the ^{1}H-NMR integrals defined for the twenty non-normalized spectra. The integral region #30 originates many cross-correlations (circled with a black line), mostly unexpected.* B) *Data matrix generated from the 2D-TOCSY spectrum. Cross-peak are due to scalar magnetic coupling between signals and are represented by black dots.*

While the 1D proton spectrum provides, through the chemical shift value, information on the chemical nature of all metabolites present in the mixture, the 2D-TOCSY provides also information on which are the signals belonging to the protons of the same molecule. The two-dimensional spectrum can resolve signals of one molecule even though they are overlapped in the 1D spectrum with others belonging to another molecule of the mixture.

Differently from the 2D-TOCSY, which gives only true correlations among signals of the same metabolite, the STOCSY autocorrelation matrix originating from the non-normalized spectra shows also false correlations between signals correctly assigned to different molecules, even with high R^2 values (Figure 1A). For example, signal #30 assigned to valine correlates only with the signals 27 and 17 (having chemical shift 2.09 and 3.44 ppm, respectively), in the 2D-TOCSY (Figure 1B), whilst the same signal correlates with a large number of integral regions, in the non-normalized autocorrelation matrix (Figure 1A): 7 (black – very good correlation), 28, 27, 26, 21, 17, 16, 6 and 4 (dark grey – good correlation), 19 and 20 (medium grey – acceptable correlation) and with signals 18 and 3 (light grey – low correlation).

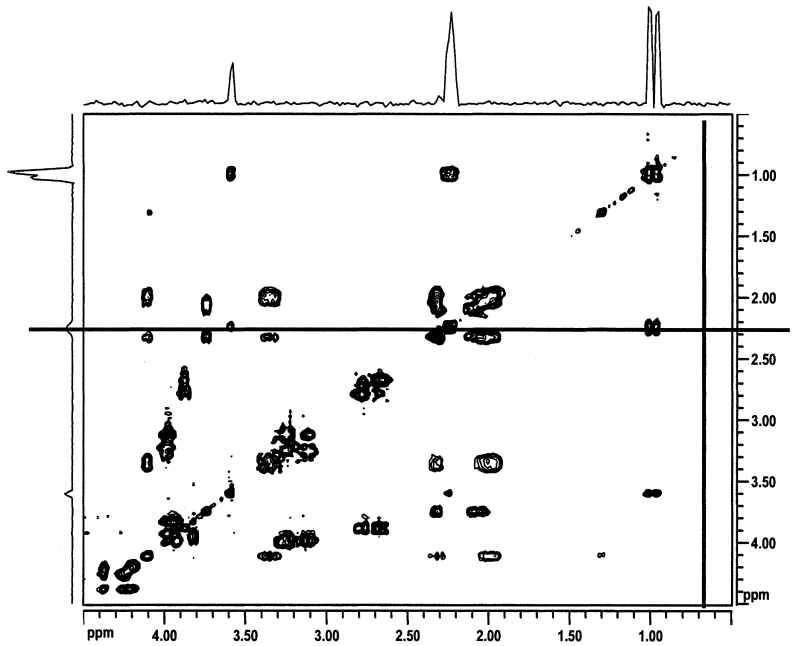

Figure 2 *2D-TOCSY spectrum acquired on the most concentrated mixture. Valine correlations are traced with two lines crossing in correspondence of the off-diagonal peak between the signal at 0,990 ppm assigned to γ,γ'-metyl protons and the signal at 2.099 ppm from β-CH*

In order to verify that the autocorrelation matrix calculated by the mathematical regression would be more similar to the spectroscopic 2D-TOCSY matrix, if the spectra were correctly normalized, each of all 20 spectra has been multiplied by the corresponding known dilution factor K before proceeding with the calculation of the autocorrelation matrix. In this way, the autocorrelation matrix shown in Figure 3A is obtained. By comparing this matrix with the one obtained by using non-normalized spectra (Figure 1A), it is possible to appreciate how the normalizing operation reduces the number of false correlations among integrals.

Except for integrals 27 and 17 (true correlations) all the others are due to the artefacts derived by a wrong dilution operation occurring during samples preparation.

In this case, only one correlation with the integral region #17 was obtained for integral region #30. Thus, the new matrix is almost characterized by the presence of only true correlations as in the 2D-TOCSY. Such a good result was achieved because the dilution factors were previously known since they were intentionally obtained during samples preparation. In true systems, the dilution factors for a correct normalization are unavailable and a proper way to calculate them is desirable.

In order to calculate the actual dilution K-factors, an algorithm has been developed in the R-project environment, namely the "TOCSY-filter algorithm". This algorithm has the purpose to minimize the number of false correlations that may be present in a non-normalized matrix but not confirmed by the TOCSY spectrum. To apply the algorithm, three false and one true cross-correlations, all with R^2 values higher than 0.8, were selected. The algorithm minimize R^2 corresponding to the false cross-correlations and,

simultaneously, maximize R^2 corresponding to the true cross-correlation, by multiplying each spectrum by an opportune dilution K-factor. This operation is repeated several times, with different pools of false and true cross-correlations, each trial providing a K-vector consisting of 20 dilution coefficients.

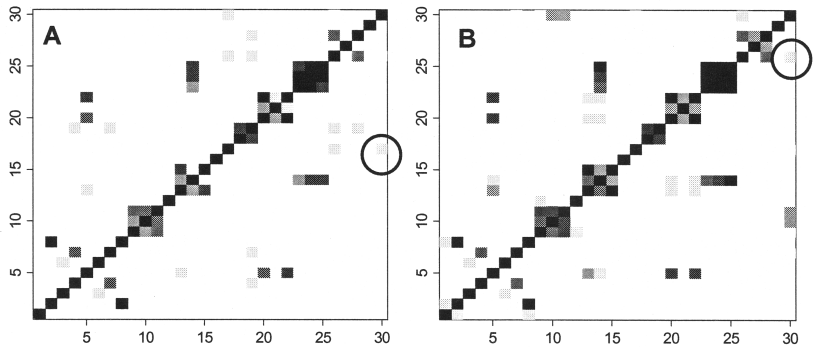

Figure 3 A) *Autocorrelation matrix obtained by multiplying the spectra of all 20 mixtures by the corresponding dilution factor K (normalization). The result is a decrease number of false correlations among integrals, as it can be seen from the number 30.* B) *Autocorrelation matrix obtained by multiplying the spectra of all 20 mixtures by the corresponding dilution factor found out by the algorithm developed in the present study.*

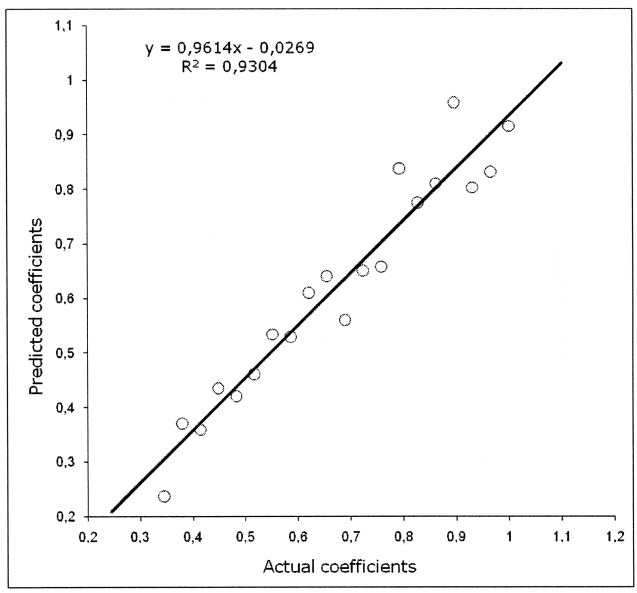

Figure 4 *Correlation between the calculated K-dilution coefficients and the actual dilution factors, that are known because artificially introduced during the preparation of the twenty mixtures.*

At the end, a mean K-vector is calculated by averaging each term over all the dilution vectors obtained by the different trials. Before calculating the mean calculated dilution coefficients, all vectors are preliminarily normalized by dividing all terms in the K-vector by the highest internal K value. In Figure 4 all the calculated K-coefficients (on the average of 6 trials) were compared with the actual ones. The result is a good linear correlation among the actual and the calculated K-factors.

The efficacy of the TOCSY-filter, is readily appreciable by looking at the new autocorrelation matrix (Figure 3B) which is obtained by multiplying the 20 raw spectra by the corresponding K-factors calculated by the algorithm.

Comparison among Figures 1A, 3A and 3B, demonstrates that the algorithm improves the reliability of the autocorrelation matrix. In fact, by considering once again the signal #30, it is possible to find the correlation dots with signals #17 and #27, that are considered as true correlations on the basis of the 2D-TOCSY spectrum.

The algorithm, in some cases, has also shown the capability to work with overlapped signals. For example, in the integral region #13 two signals are present (13 and 13'): at 3.835 ppm for histidine and at 3.804 ppm for serine (see Table 1). The serine correlates with signal #15 at 3.679 ppm, whilst histidine is connected with signal #21 at 3.037 ppm and signal #22 at 2.959 ppm. The algorithm did not fail to find correlations between integral region #13 and signal #15 for serine and between integral region #13 and signal #22 for histidine.

2.5.2 Application of the normalization algorithm on a real food system. A spectral dataset processed with a wrong normalization method produces unreliable statistical results, since the meaningful variance of the metabolite levels among all samples may be hindered by large differences of solvent quantity. Such a variability of the solutes/solvent proportions among the samples determines also the poor co-variance between the observables of a molecule when measured by different analytical techniques on the same samples set. The co-variance analysis is the basis of the statistical hetero-spectroscopy (SHY), which is an extension of STOCSY.[23] In principle, a co-variance matrix between NMR and HPLC data will show correlations among the chemical shift of NMR signals and the retention times of the corresponding metabolite in the chromatogram (Figure 5).

For the biological extracts, the co-variance matrix has a large number of cross-correlations between NMR signals and chromatographic peaks (Figure 5). The number is however much higher than the expected and, as for STOCSY, this is diagnostic for the presence of false correlations associated with a wrong normalization procedure. In the co-variance matrix, false correlations are due to different dilutions of the samples, like those generated by different amounts of residual water in the lyophilized samples' powders. Such variations may also be induced by some inevitable sampling errors, especially in case of non homogeneous food matrices. This is the case of freeze-dried cherry tomatoes powders, where it is possible to collect, by chance, different anatomic parts of the fruit, such as seeds, peel or pulp with different contents of extractable matter.

Consequently, in order to minimize the dilution effect on NMR spectra, the TOCSY-filter algorithm has been applied on the ^{1}H-NMR autocorrelation matrix obtained from spectra acquired on 11 cherry tomatoes extracts. The resulting STOCSY matrix (Figure 6A) has been compared with a 2D ^{1}H-TOCSY spectrum acquired on a single chloroform extract of cherry tomato (Figure 6B), in order to identify true and false cross-peaks in the autocorrelation matrix.

The comparison of the two matrices points out the presence of false cross peaks, for example those circled with a red line in Figure 7. The application of the TOCSY-filter algorithm minimizes the presence of false cross peaks in the autocorrelation matrix (data not shown), by applying 11 dilution K-factors, one for each mixture.

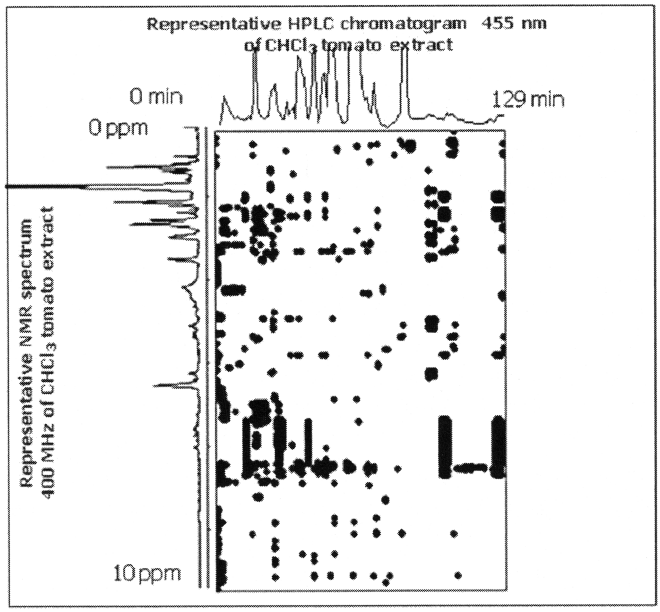

Figure 5 *HPLC-NMR co-variance matrix calculated for non-normalized spectra and chromatograms of natural samples, consisting of chloroform extracts from cherry tomatoes*

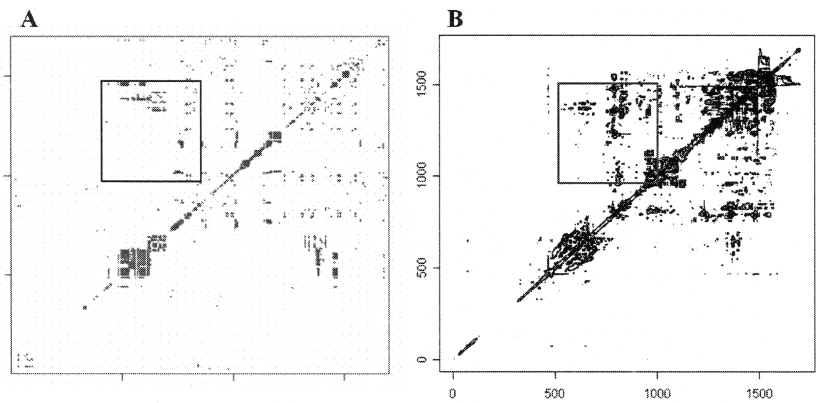

Figure 6 A) *Non-normalized autocorrelation array ^{1}H-NMR (1700 x 1700) showing cross-correlations with R^2> 0.8.* **B)** *2D-^{1}H-TOCSY, acquired at 800 MHz, on a cherry tomato sample chosen .*

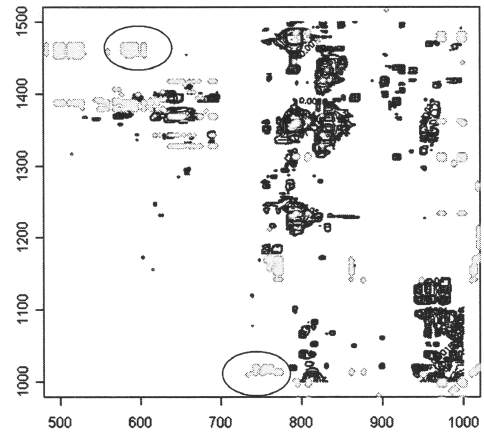

Figure 7 *Superimposition between non-normalized autocorrelation array with only true correlation, in grey, and ^{1}H-TOCSY acquired on a cherry tomato sample, in black. Some false cross peaks are circled.*

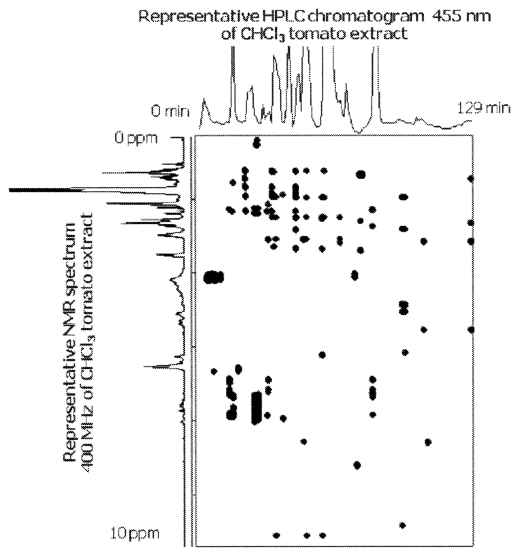

Figure 8 SHY NMR-HPLC *co-variance matrix between the ^{1}H-NMR signals and the chromatogram recorded at 455 nm on the same sample set, showing a minimized number of false cross peaks. The optimization has been performed by applying the TOCSY-filter algorithm.*

Once the dilution K-factors are calculated, they have been used as normalization coefficients for the NMR spectra, as well as for all chromatograms. The resulting SHY co-variance matrix shows the capability of the algorithm to clean out the false cross-correlation, thus improving the reliability of a non-hyphenated HPLC-NMR coupling, that

can be used to assign a structure to a molecule whose peak is eluted at the corresponding retention time (Figure 8). It is worth noting that the TOCSY-filter algorithm minimizes the R^2 of artefacts whilst does not affect or maximizes the regression coefficients of true cross-correlations. Thus, the filtering out of a large number of "signals" in the 2D covariance matrix, is not corresponding to the simple cut off procedure obtained by increasing the threshold for the elements to be shown in the image.

Although several resonances are readily observed between 6 and 7 ppm, consistent with the presence of lycopenes in the tomato extracts, not all the main chromatographic peaks, neither all the main NMR signals, shows co-variance between the two techniques. This is easily explained by the assumption that not all the peaks and the signals are resolved, and the overlap among signals and, more frequent, among peaks, is critical for the obtainment of clear covariance. The enhancement of the SHY matrix relies on the improvement of the HPLC separation and of the NMR resolution. However, whilst the latter parameter is constant for a given magnetic field, and can be increased at a larger cost, the first technique can be easily improved by changing the elution gradients and the column performances.

3 CONCLUSION

The TOCSY-filter algorithm, at first developed on a training dataset consisting of NMR spectra acquired on 20 mixtures of 10 different metabolites, has been shown to be able to correct the dilution errors that have been artificially and purposely introduced during sample preparation. Dilution errors is proven to lead to an incorrect normalization of the proton spectra, which is a necessary step prior the multivariate analysis.

The algorithm requires the acquisition of a two-dimensional ^{1}H- TOCSY spectrum on a single sample, chosen among the whole set that will undergo the chemometric analysis. The TOCSY-filter is a simple algorithm, suitable for a future inclusion in the standard procedures so far adopted in metabonomics, as a preliminary step before statistical analysis. The resulting elimination of the effects of incorrect dilution procedures as source of errors, determines an improvement in the discovery of the biological source of variance, which is the main goal in metabonomics, including the applications in foodomics science.

In this field, further information can be gained from covariance analysis of data obtained by two different techniques, either spectroscopic and chromatographic, on the same sample set, in the so called SHY applications. However, when dilution errors are larger than, or comparable to, the biological variance of metabolites concentrations, the co-variance information is poor and the SHY approach is vain. The elimination, or reduction, of the dilution errors, obtained by application of the TOCSY-filter algorithm, may improve the SHY applications, rendering the approach a valid alternative to the hyphenated HPLC-NMR analysis.

References

1 Fiehn, O. *Plant Mol. Biol.* 2002, **48**, 155-171.
2 R.N. Trethewey, A.J. Krotzky and L. Willmitzer, 1999, *Curr. Opin. Plant Biol.*, **2**, 83.
3 J.C. Lindon and J.K. Nicholson, *Ann. Rev. Anal. Chem.*, 2008, **1**, 45.
4 W.M.T. Fan, 1996, *Prog. Nucl. Magn. Reson. Spectrosc.*, **28**, 161.
5 E.M. Lenz and I.D. Wilson, 2007, *J. Proteome Res.*, **6**, 443.
6 J.L. Wolfender, S. Rodriguez and K. Hostettmann, 1998, *J. Chromatogr. A*, **794**, 299.
7 H. Dai, C. Xiao, H. Liu and H. Tang, 2010, *J. Proteome Res.*, **9**, 1460.
8 J. K Nicholson and J. C. Lindon, 2008, *Nature*, **455**, 1054.
9 E. Holmes and H. Antti, 2002, *Analyst*,**127**,1549

10 A.M. Weljie, J. Newton, P. Mercier, E. Carlson and C.M. Slupsky, 2006, *Anal. Chem.,* **78** , 4430.
11 D.S. Wishart, 2008, *Trends Food Sci. Technol.*, **19**, 482.
12 L. Tarachiwin, K. Ute, A. Kobayashi and E. Fukusakii, 2007, *J. Agric. Food Chem.*, **55**, 9339.
13 C.N. Xiao, H. Dai, H.B. Liu,Y.L. Wang and H.R. Tang, 2008, *J. Agric. Food Chem.*, **56**, 10142.
14 J.A. Donarski, S.A. Jones and A.J. Charlton, 2008, *J. Agric. Food Chem.*, **56**, 5451.
15 H.S. Son, K.M. Kim, F. Van den Berg, G.S. Hwang, W.M. Park, C.H. Lee and Y.S. Hong, 2008, *J. Agric. Food Chem.*, **56**, 8007.
16 G.E. Pereira, J.P. Gaudillere, C. Van Leeuwen, G. Hilbert, M. Maucourt, C. Deborde, A. Moing and D. Rolin, 2006, *Anal. Chim. Acta*, **563**, 346.
17 A.P. Sobolev, A.Segre and R. Lamanna, 2003, *Magn. Reson. Chem.,* 41, 237.
18 H.P.J.M. Noteborn, A. Lommen, R.C. Van der Jagt and J.M. Weseman, 2000, *J. Biotech.,* **77**, 103.
19 R. Goodacre, D. Broadhurst, A.K. Smilde, B.S. Kristal, J.D. Baker, R. Beger, C. Bessant, S. Connor, G. Capuani, A. Craig, T. Ebbels, D. B. Kell, C. Manetti, J. Newton, G. Paternostro, R. Somorjai, M. Sjöström, J. Trygg, F. Wulfert, 2007, *Metabolomics*, 3, 231.
20 A. Craig, O. Cloarec, E. Holmes, J.K. Nicholson and J.C. Lindon, 2006, *Anal. Chem.*, **78**, 2262.
21 F. Dieterle, A. Ross, G. Schlotterbeck and H. Senn, 2006, *Anal. Chem.*, **78**, 4281.
22 O. Cloarec, M.E. Dumas, A. Craig, R.H. Barton, J. Trygg, J. Hudson, C. Blancher, D. Gauguier, J.C. Lindon, E. Holmes and J. Nicholson, 2005, *Anal. Chem.*, **77**, 1282.
23 D.J. Crockford, E. Holmes, J.C. Lindon, R.S. Plumb, S. Zirah, S.J. Bruce, P. Rainville, C.L. Stumpf and J.K. Nicholson, 2006, *Anal. Chem.*, **78**, 363.